Fair Principal Component Analysis for Correcting Population Stratification

Joseph Valencia
School of Electrical Engineering and Computer Science
Oregon State University
Corvallis, OR
valejose@oregonstate.edu

Abstract

Principal component analysis (PCA) is a common technique for dimensionality reduction. A growing awareness of algorithmic bias in recent years has motivated the development of unsupervised methods that incorporate fairness objectives. Kamani et al. [2020] introduces an efficient gradient-descent based approximation of PCA that preserves Pareto optimal reconstruction loss between user-defined subgroups. Similarly, Tantipongpipat et al. [2020] describes several methods for optimizing various fairness objectives, including Nash social welfare. An important potential application for such methods is in genome-wide association studies, where principal components are commonly used to control for the effect of ancestral population stratification as introduced in Price et al. [2006]. I re-implement both these algorithms along with conventional PCA and compare their performance on the Adult and 1000 Genomes datasets. I also use simulated single nucleotide polymorphism data to test the downstream consequences of fairness objectives on GWAS studies.

1 Introduction

In the era of big data, dimensionality reduction techniques enable researchers to perform exploratory analysis of high-dimensional data and reduce the complexity of machine learning models. Principal components analysis (PCA) is among the most ubiquitous methods for dimensionality reduction for its desirable theoretical properties and ease of use. PCA is used in nearly every application domain [Jolliffe and Cadima, 2016].

Algorithmic bias refers to the tendency of machine learning and other automated decision making tools to capture and propagate the societal inequities embedded in training data. As increasing amounts of the economy and society are automated, machine learning models are used to recommend parole decisions, filter resumes, allocate loans, and much more [Corbett-Davies et al., 2017; Raghavan et al., 2020]. These applications have the potential to reinforce discrimination based on protected characteristics such as race and gender, particularly when deep learning is involved, as the algorithms are very hard to scrutinize. For example, datasets for facial recognition have been found to be biased against individuals with darker skin and women [Buolamwini and Gebru] . The majority of work in reducing algorithmic bias has focused on supervised methods like classification, but fairness is also relevant for unsupervised learning and representation learning. One line of work has been to obfuscate sensitive attributes in the latent space that is learned by PCA [Olfat and Aswani, 2018]. Another is to ensure that the learned representations of subgroups fairly capture all subgroups, which is the focus of this paper.

An under-explored area of algorithmic fairness is in the field of computational biology. Race is a socially constructed abstraction rather than a scientifically meaningful attribute, but geneticists

Preprint. Under review.

genetic variation and historically subpopulations. In biomedical applications, it is typically not desirable to obscure ancestral information, as there can be significant associations between heredity and disease, and treatments However, fairness The vast majority of genome-wide association (GWAS) studies have been on populations of predominantly European descent Gaspar and Breen [2019]. As thesse studies expand to incorporate other populations, so genetic diseases that disproportionately affect other populations could be understudied relative to those commonly found among those of European ancestry.

In this paper, I implement two modified versions of principal component analysis that optimize different fairness objectives. The first is a Frank-Wolfe approximation of a semidefinite program for maximizing the Nash social welfare metric between sensitive groups [Tantipongpipat et al., 2020]. The second is a gradient descent method that preserves Pareto efficient reconstruction fidelity among groups Kamani et al. [2020]. I replicate experiments on the Adult and Credit datasets that are commonly used to assess fairness-related algorithms. Further, I apply these two algorithms to a new dataset, the 1000 Genomes dataset and replicate experiments from [Gaspar and Breen, 2019].

2 PCA Algorithms

Principal components analysis aims to project high-dimensional data $X \in \mathbb{R}^{m \times n}$ into a lower-dimensional latent space $X_d \in \mathbb{R}^{m \times d} \quad d < n$ that maximizes the variation of the dataset. The canonical solution for PCA comes from the eigendecomposition of the correlation matrix of the data.

$$V \Lambda V^T = X^T X \tag{1}$$

Where V is orthonormal and Λ is a diagonal matrix con PCA solution is thus $= V_d$ is a matrix whose rows are the eigenvectors corresponding to the d leading eigenvalues. PCA can also be obtained using the singular value decomposition of X.

From an optimization perspective, PCA can be viewed as the solution to the problem

$$\begin{aligned} \text{PCA}_d \quad &= \arg\min_A \|X - XAA^T\|_F \\ \text{s.t.} \quad & A \in \mathbb{R}^{n \times d} \\ & A^T A = \mathbb{I}_d \end{aligned}$$

This is non-convex problem in A, so the spectral method outlined above is usually preferred.

2.1 Fairness Objectives

The ability to remedy algorithmic unfairness relies a define fairness in an ethical sense and to represent these values in mathematical form. There is an important distinction between individual fairness and group fairness. Because conventional PCA maximizes overall variance of the dataset in the latent space, it can be said to maximize average-case individual fairness. However, this could mean compromising reconstruction fidelity when aggregated at the level of important sub-populations.

The two definitions of group fairness in this paper originate from welfare economics. Economists use the concept of utility to describe the gain or contentment that is received from a market transaction. Pareto efficiency refers to a situation in which no party can improve their utility u_i without reducing the utility of another party.

$U = [u_1 \quad u_2 \quad \cdots \quad u_n]$ is Pareto optimal iff $\nexists \quad \tilde{U}$ such that $\tilde{u}_i \leq u_i \quad \forall i = 1 \cdots n$ and $\tilde{u}_i < u_i$ for some i.

Nash social welfare emerged from John Nash's work on. Intuitively, it is the product of individual utilities.

$$NSW = \Pi_i^n u_i \tag{2}$$

In either case, we can consider the utility of each group to be a function of its PCA loss.

2.2 Pareto fair PCA

The method of Kamani et al. [2020], which I refer to as Pareto-PCA is a. This is a multi-objective optimization problem, and the Pareto optimality condition requires that each update state be non-decreasing for each objective.

At each time step t, the gradient for the overall PCA loss is

$$\nabla_{A_t} = -2X^T X A_t$$

And the gradients for each fairness objective is

$$\nabla_{\phi(L(A_t)_i)} = -2\frac{\partial \phi(A_t)_i}{\partial A_t} X_i^T X_i A_t$$

Where X_i is the data matrix for sensitive group $i \in [k]$ and $\phi(L(A_t)_i)$ is a positive nonlinear function of the PCA reconstruction loss for group i. The dual problem given in the appendix of the original work reduces to a quadratic program

$$\begin{aligned} \max_{\lambda} \quad & \lambda^T G^T G \lambda \\ \text{s.t.} \quad & 1^T \lambda = 1 \\ & \lambda \succeq 0 \end{aligned}$$

Where G is a matrix whose rows are the vectorized gradient of each objective. The resulting probability vector λ is used to weight the gradients, which identifies the Pareto optimal descent direction. After each gradient descent step, the running we obtain UV^T, from the singular value decomposition $A_{t+1} = U\Sigma V^T$, which is the solution to the orthogonal Procrustes problem. This enforces the orthonormal property of $A^T A$ as required by PCA

2.3 Nash social welfare PCA

The method of Tantipongpipat et al. [2020], which I refer to as Frank-Wolfe-Nash, solves the optimization

$$\begin{aligned} \max_{Q \in \mathbb{R}^{n \times n}} \quad & \sum_{i=1}^{k} \log(tr(X_i X_i^T Q) + \alpha \| X_i^T X_i \|_F) \\ \text{s.t.} \quad & tr(Q) = d \\ & 0 \preceq Q \preceq 1 \end{aligned}$$

Where X_i is defined in the preceding section, and $Q = AA^T$ where A refers to the principal component matrix. The second term in the objective function serves as a regularization.

The VV^T provides a linear maximizer that enables a Frank-Wolfe algorithm.

3 Correcting for population stratification

The motivation for this study is the prevalence of principal components analysis in genome-wide association studies. The unit of observation in a GWAS is called a single-nucleotide polymorphism (SNP), which is the loci within the genome that correspond to most of the variation between individuals. The alleles contained at a SNP is called the genotype is compared at tens of thousands of SNPs to . When population stratification .These studies involve comparing genotype-phenotype associations, i.e. the statistical prevalence of particular alleles and disease states. The control group is the reference population without a disease or trait, whereas the case group. It can be difficult to ensure exact ancestral comparability of these two groups, and so each SNP is regressed against the identified PCs as described in [Price et al., 2006], which removes these spurious associations.

4 Experiments

4.1 Datasets

4.2 PCA performance

I trained the vanilla PCA method (eigendecomposition) and the Frank-White-Nash (FWN) algorithm on both the Adult and 1000 Genomes datasets and monitored their performance in reconstruction loss and Nash social welfare. For the 1000 Genomes trials, calculating the eigenvalues of the 5000×5000 correlation matrix caused my program to crash due to a multiprocessing error, so I had to fall back to the scikit-learn implementation of PCA, which uses a more efficient solver. I was unable to complete my implementation of Pareto-PCA method because I could not identify a normalization and regularization strategy that simultaneously 1) places the gradients for each objective on an equivalent scale, which is a necessary step for solving the dual problem for the descent direction 2) permits the descent direction to asymptotically vanish to zero, which is necessary for the stopping condition.

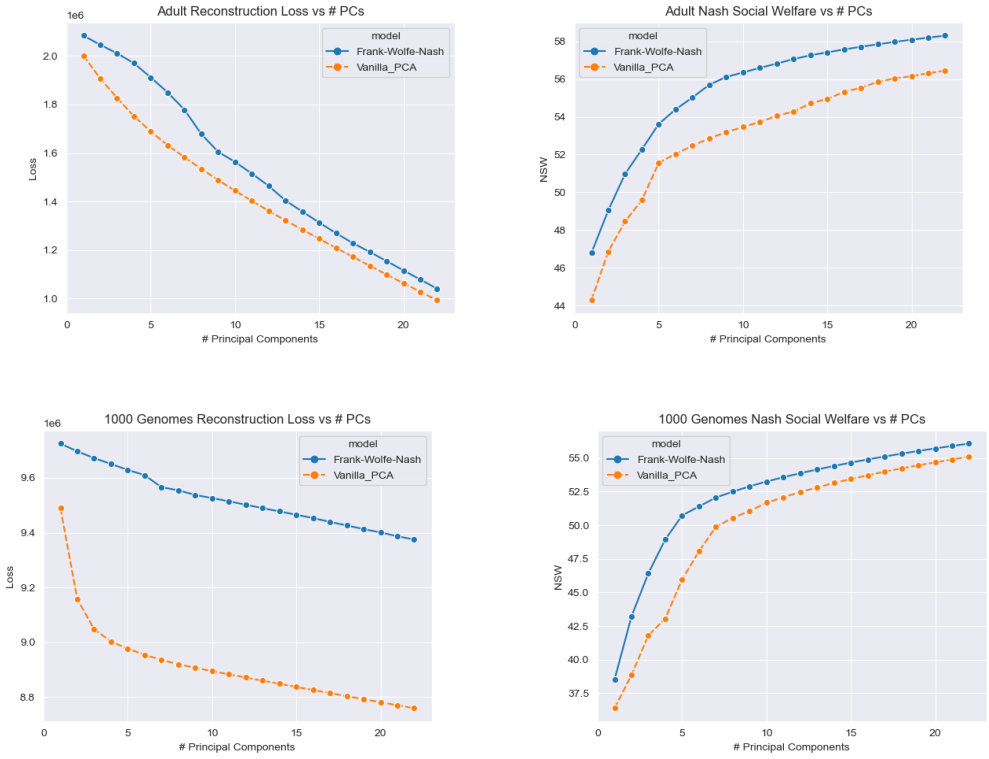

Figure 1: Three simple graphs

On the Adult dataset, as shown in Figure **??**, the Frank-White-Nash achieves superior performance on the Nash social welfare metric. Its reconstruction loss is higher than that of Vanilla PCA, although this gap narrow substantially as more principal components are included. On the 1000 Genomes, the gap in reconstruction loss is much higher, with additional principal components leading to linear improvements in the FWN model, while the first few principal components lead to a sublinear improvement in the vanilla PCA model. As expected, the FNW still performs better on the Nash objective for the 1000 Genomes dataset.

4.3 Classification

I reproduce the experiments in Gaspar and Breen [2019] , which applies PCA to the 1000 Genomes in order to assess how the reduced dimensions capture specific population clusters. I plot the first two principal components from both methods in Figure **??**. AFR refers to African, AMR: Admixed

American ,EAS :East Asian, EUR European: SAS South Asian. It appears that the latent space learned by Frank-Wolfe-Nash retains more of the variation within the East Asian and Admixed American superpopulation group while compromising the visual separation of many of the groups, particularly the African superpopulation. Admixed refers to the fact that the AMR group contains individuals from all across the Americas, so it makes sense that genotypes from this sample are among the most heterogenous.

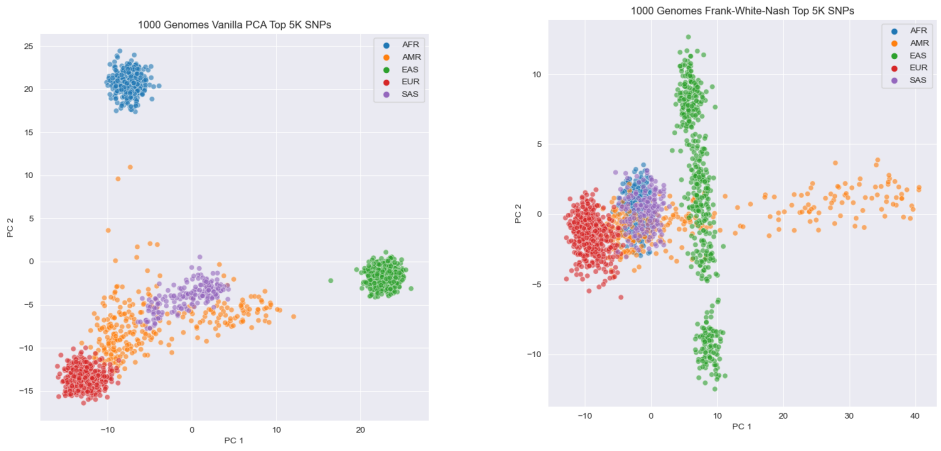

Figure 2: Three simple graphs

Next, I assess the ability of both PCA methods to recover the true demographic labels provided by the 1000 Genomes set. I use five-fold cross-validation and repeat the analysis ten times, aggregating scores over each one. I use two different methods of recovering population status, by training a K-Nearest Neighbors classifier on the first two PCs (i.e. the two dimensions in Figure **??**. This roughly captures the visual separation in the scatter plot. As a higher-powered model, I train a support vector machine classifier on the first ten PCs. I report the result of these analyses on both the broad population families **??** as well as the more fine-grained populations **??**. Note that in the latter case, the principal components for Frank-Wolfe-Nash have only been learned using the broader groupings. I hoped that the superior within-family variation would enable the dimensions learned through this strategy to better capture the more subtle intra-family distinctions.

	F1			
	SVM (10 PCs)		KNN (2 PCs)	
Super Population	Vanilla PCA	Frank-Wolfe-Nash	Vanilla PCA	Frank-Wolfe-Nash
AFR	0.998+-0.0	0.991+-0.001	0.998+-0.0	0.821+-0.003
AMR	0.982+-0.002	0.924+-0.004	0.882+-0.004	0.638+-0.007
EAS	1.0+-0.0	1.0+-0.0	1.0+-0.0	0.977+-0.001
EUR	0.992+-0.001	0.981+-0.001	0.978+-0.001	0.953+-0.002
SAS	1.0+-0.0	0.93+-0.005	0.897+-0.004	0.392+-0.011

5 Conclusion and future work

The fairness-aware versions of PCA discussed in this paper appear to reduce the usefulness of principal components in visually separating genotypes by population. It appears that population stratification is mostly a function of inter-group variance, which is compromised by these methods in favor of preserving intra-group fairness. However, it is likely that there are applications for which this matters more, particularly when other populations besides Europeans are considered. Future

	F1			
	SVM (10 PCs)		KNN (2 PCs)	
	Vanilla PCA	FW-Nash	Vanilla PCA	FW-Nash
Population / Location				
Bengali / Bangladesh	0.944+-0.006	0.953+-0.006	0.883+-0.006	0.53+-0.01
Chinese Dai / Xishuangbanna, China	0.767+-0.011	0.718+-0.011	0.302+-0.011	0.702+-0.01
Colombians / Medellin, Colombia	0.853+-0.008	0.848+-0.009	0.581+-0.014	0.407+-0.013
Esan / Nigeria	0.592+-0.013	0.609+-0.011	0.237+-0.01	0.191+-0.01
Finnish / Finland	0.988+-0.002	0.992+-0.002	0.707+-0.01	0.614+-0.01
Gambian / Western Divisions,Gambia	0.934+-0.005	0.929+-0.005	0.279+-0.011	0.139+-0.007
Han Chinese / Beijing, China	0.792+-0.011	0.778+-0.01	0.205+-0.011	0.728+-0.009
Iberian Population / Spain	0.643+-0.011	0.674+-0.01	0.287+-0.009	0.328+-0.011
Japanese / Tokyo, Japan	0.995+-0.001	0.998+-0.001	0.211+-0.011	0.997+-0.001
Kinh / Ho Chi Minh City, Vietnam	0.776+-0.01	0.675+-0.014	0.261+-0.013	0.639+-0.013
Luhya / Webuye, Kenya	1.0+-0.0	1.0+-0.0	0.48+-0.01	0.152+-0.011
Mende / Sierra Leone	0.904+-0.007	0.883+-0.007	0.147+-0.014	0.073+-0.009
Northern/Western European Ancestry	0.967+-0.002	0.945+-0.003	0.571+-0.006	0.543+-0.008
Peruvians / Lima, Peru	0.96+-0.004	0.962+-0.004	0.872+-0.007	0.964+-0.004
Puerto Ricans / Puerto Rico	0.889+-0.005	0.817+-0.007	0.816+-0.007	0.354+-0.013
Punjabi / Lahore, Pakistan	0.947+-0.005	0.864+-0.009	0.787+-0.006	0.054+-0.008
Southern Han Chinese	0.83+-0.007	0.795+-0.008	0.329+-0.009	0.763+-0.007
Toscani / Italy	0.713+-0.008	0.743+-0.008	0.272+-0.012	0.195+-0.012
Yoruba / Ibadan, Nigeria	0.698+-0.008	0.659+-0.008	0.2+-0.011	0.1+-0.009

work will seek to apply these fairness methods to the downstream GWAS task and to speed up the efficiency for large SNP arrays.

Acknowledgments and Disclosure of Funding

I would like to thank Professor Xiao Fu for his capable instruction this term, as well as the authors of the papers on which this work is based.

References

Mohammad Mahdi Kamani, Farzin Haddadpour, Rana Forsati, and Mehrdad Mahdavi. Efficient Fair Principal Component Analysis. *arXiv:1911.04931 [cs, math, stat]*, March 2020. URL `http://arxiv.org/abs/1911.04931`. arXiv: 1911.04931 version: 2.

Uthaipon Tantipongpipat, Samira Samadi, Mohit Singh, Jamie Morgenstern, and Santosh Vempala. Multi-Criteria Dimensionality Reduction with Applications to Fairness. *arXiv:1902.11281 [cs, math]*, June 2020. URL `http://arxiv.org/abs/1902.11281`. arXiv: 1902.11281.

Alkes L. Price, Nick J. Patterson, Robert M. Plenge, Michael E. Weinblatt, Nancy A. Shadick, and David Reich. Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics*, 38(8):904–909, August 2006. ISSN 1546-1718. doi: 10.1038/ng1847. URL `https://www.nature.com/articles/ng1847`. Number: 8 Publisher: Nature Publishing Group.

Ian T. Jolliffe and Jorge Cadima. Principal component analysis: a review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374(2065):20150202, April 2016. ISSN 1364-503X, 1471-2962. doi: 10.1098/rsta.2015.0202. URL `https://royalsocietypublishing.org/doi/10.1098/rsta.2015.0202`.

Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. Algorithmic Decision Making and the Cost of Fairness. page 10, 2017.

Manish Raghavan, Solon Barocas, Jon Kleinberg, and Karen Levy. Mitigating bias in algorithmic hiring: evaluating claims and practices. In *Proceedings of the 2020 Conference on Fairness,*

Accountability, and Transparency, pages 469–481, Barcelona Spain, January 2020. ACM. ISBN 978-1-4503-6936-7. doi: 10.1145/3351095.3372828. URL https://dl.acm.org/doi/10.1145/3351095.3372828.

Joy Buolamwini and Timnit Gebru. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. page 15.

Matt Olfat and Anil Aswani. Convex Formulations for Fair Principal Component Analysis. *arXiv:1802.03765 [cs, math, stat]*, November 2018. URL http://arxiv.org/abs/1802.03765. arXiv: 1802.03765.

Héléna A. Gaspar and Gerome Breen. Probabilistic ancestry maps: a method to assess and visualize population substructures in genetics. *BMC Bioinformatics*, 20(1):116, March 2019. ISSN 1471-2105. doi: 10.1186/s12859-019-2680-1. URL https://doi.org/10.1186/s12859-019-2680-1.